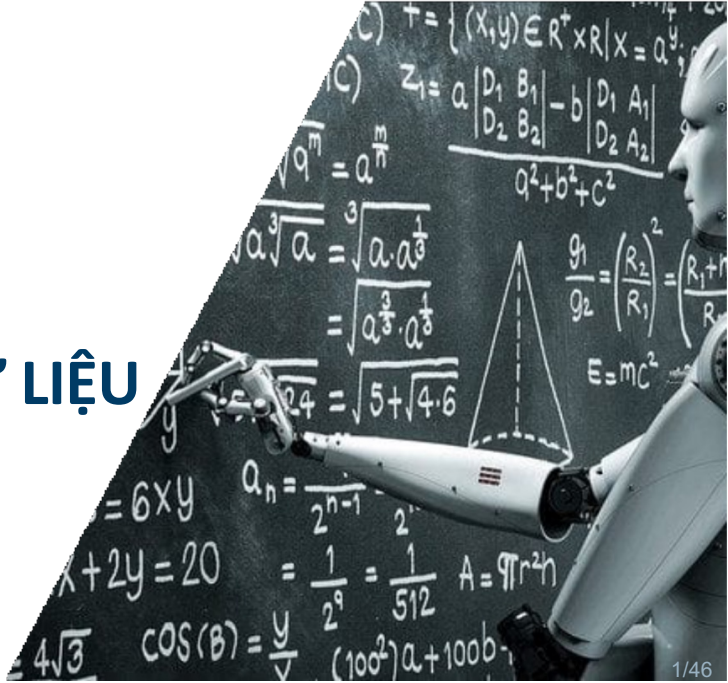
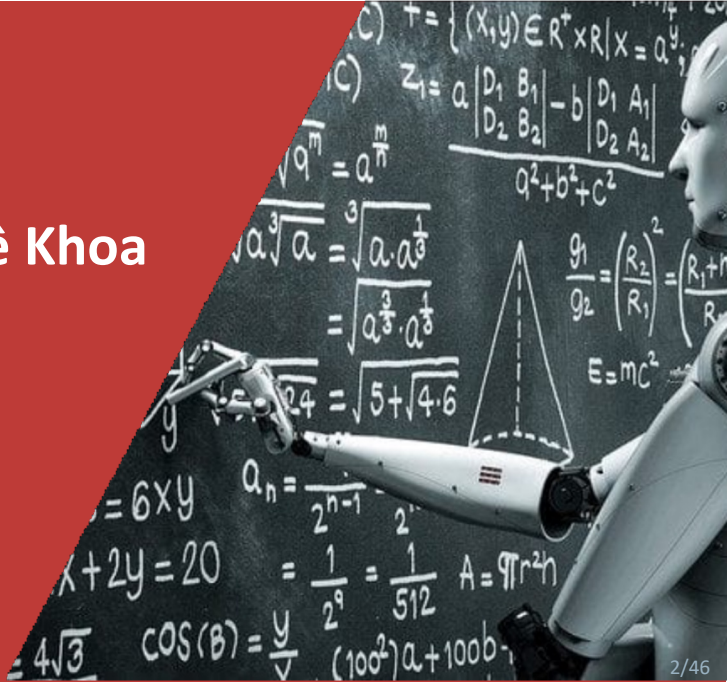


# NHẬP MÔN KHOA HỌC DỮ LIỆU

TS. NGUYỄN HUYỀN CHÂU  
AI LAB - TLU



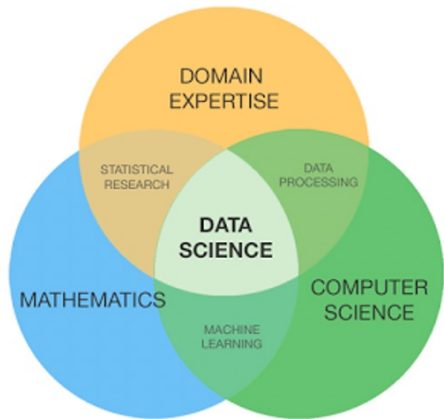


## Khoa học dữ liệu là gì

Nội dung khoá học

Thu thập và đọc dữ liệu

# Khoa học dữ liệu là gì?



***Khoa học liên ngành về:***

- *Phương pháp*
- *Quy trình*
- *Công cụ*

***Để hình thành tri thức từ dữ liệu***

# Tại sao bây giờ?

## History of data science

### Corporate, Industry & Business People



1955s

Taylorism was adopted majorly across industrial revolution



1965s

Peter Drucker introduced the idea - knowledge worker



1975s

Kaizen and Toyota Production System is being adopted in manufacturing



1985s

Excel became the go to tool among knowledge workers



1995s

Internet makes the world connected



2000s

DOTCom Boom - every company take their business online



2005s

Service industry transformed customer experience and they improved with data



2010s

Smartphones intensified customer obsession and era of on-demand economy starts



2015s

Machine Learning and AI becomes mainstream. Computation intensified with GPUs

### Academia, Engg, Math, Computer Hobbyists



1955s

Computers and AI were extensively researched after World War 2



1965s

The HP 9100A, the world's first desktop computer



1975s

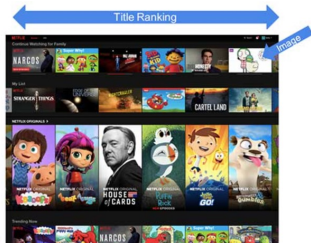
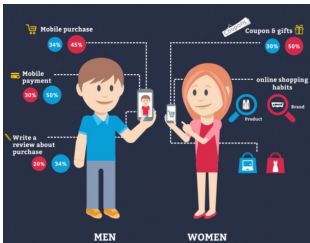
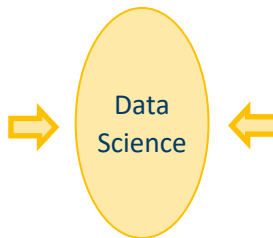
John Tukey through his book 'Exploratory Data Analysis' showed the importance of visualizing data



1985s

Operating Systems made its debut in the market with MS Windows

# Dữ liệu hoá và dân chủ hoá phân tích dữ liệu



## Các nhiệm vụ khoa học dữ liệu

- **Khảo sát thực tế:** Thu thập dữ liệu chủ động (tự tạo) hay bị động (có sẵn) để đánh giá phản ứng của thế giới, từ đó rút ra chiến thuật hoạt động tốt nhất. VD: A/B testing cho phát triển web
- **Phát hiện mẫu:** Phát hiện các mẫu và cụm tự nhiên, từ đó có thể chia-để-trị. VD: digital marketing, quảng cáo có target
- **Dự đoán sự kiện tương lai:** Thay vì phản ứng với tình thế, có thể chuẩn bị trước cho tình thế. VD: tối ưu hoá trong lập kế hoạch
- **Hiểu con người và thế giới:** Bởi vì động lực và hành vi là điểm khởi đầu của mọi hành động. VD xử lý ngôn ngữ tự nhiên, thị giác máy tính, ...

# Quy trình khoa học dữ liệu

1. Define the problem



2. Collect the data



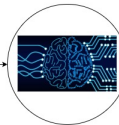
3. Preprocess the data



4. Explore the data



5. Build models



6. Make decisions





# Các nghề nghiệp khoa học dữ liệu

- **Cấp quản lý:** Giám đốc dữ liệu (CDO), kiến trúc sư dữ liệu

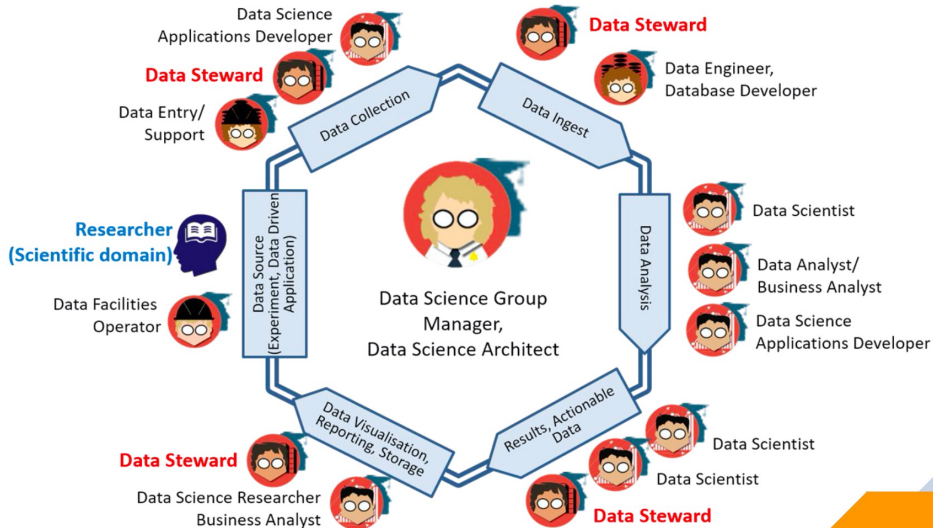
## Phân tích dữ liệu:

- **Chuyên gia DL:** Nhà khoa học, nhà thống kê, người phân tích
- **Kỹ sư:** Kỹ sư dữ liệu, kỹ sư học máy/AI/CV/NLP
- **Nhân viên:** Nhân viên phân tích kinh doanh

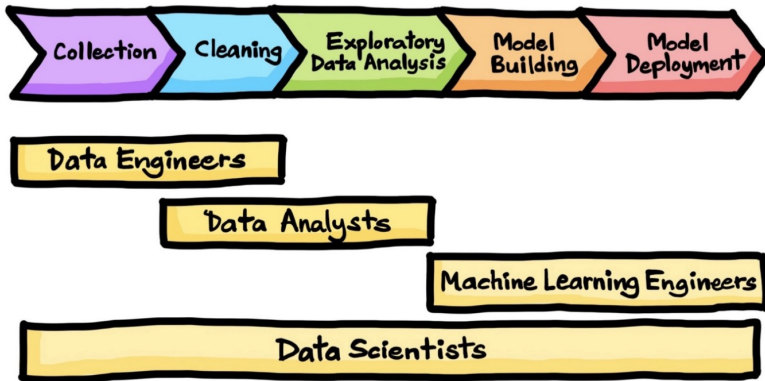
## Bảo trì dữ liệu:

- **Chuyên gia CSDL:** Quản trị viên CSDL (lớn)
- **Kỹ thuật viên:** Vận hành hệ thống CSDL (lớn)
- **Nhân viên:** Nhân viên nhập liệu

# Các nghề nghiệp và vòng đời khoa học dữ liệu



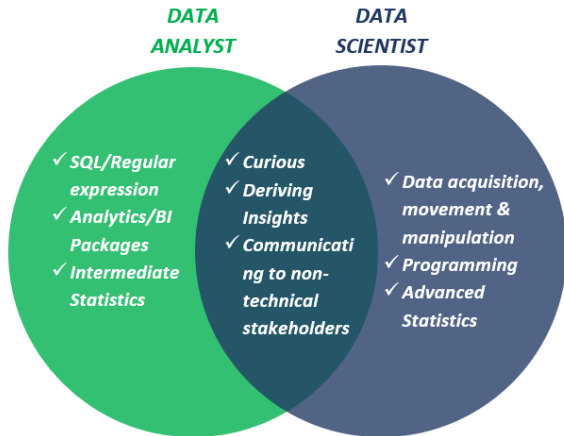
# Các nghề nghiệp và vòng đời khoa học dữ liệu





# Tập kỹ năng khoa học dữ liệu

- Thu thập dữ liệu
- Biến đổi: làm sạch, tính toán
- Trình diễn: đồ thị, bảng biểu, hình vẽ
- Tìm hiểu, phân tích dữ liệu
- Phát hiện quy luật, đặc trưng
- Xử lý dữ liệu lớn

# Tập kỹ năng cho mỗi nghề nghiệp



# Tập kỹ năng cho mỗi nghề nghiệp

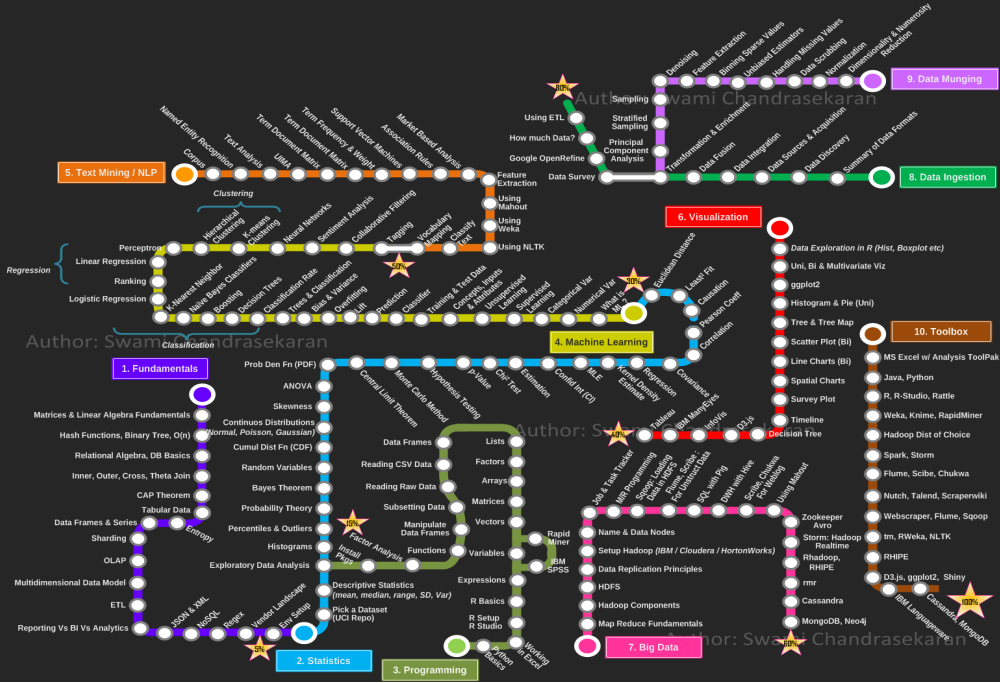
	<h2>Data scientist vs. data engineer</h2>		
DATA SCIENTIST	DATA ENGINEER		
<p><b>MAIN DUTIES</b></p> <p>Machine learning, AI algorithms, building specialized models, maintaining clean data sets</p>	<p><b>MAIN DUTIES</b></p> <p>Software development, building and maintaining data pipelines, maintaining databases and processing systems</p>		
<p><b>MAIN PROGRAMMING LANGUAGES</b></p> <p>R or Python</p>	<p><b>MAIN PROGRAMMING LANGUAGES</b></p> <p>Java and Python</p>		
<p><b>MAIN TOOLS</b></p> <p>MapReduce, Hadoop, Hive, Spark, Gurobi Optimizer, MySQL</p>	<p><b>MAIN TOOLS</b></p> <p>Hadoop, NoSQL databases, Spark, relational database management systems</p>		
<p><b>OTHER SKILLS</b></p> <p>Interpersonal communications, team building, boardroom presences</p>	<p><b>OTHER SKILLS</b></p> <p>Team building, team-oriented, comfort switching between technologies</p>		

# Tập kỹ năng cho mỗi nghề nghiệp

## Data Scientist vs. Machine Learning Engineer vs. Data Engineer

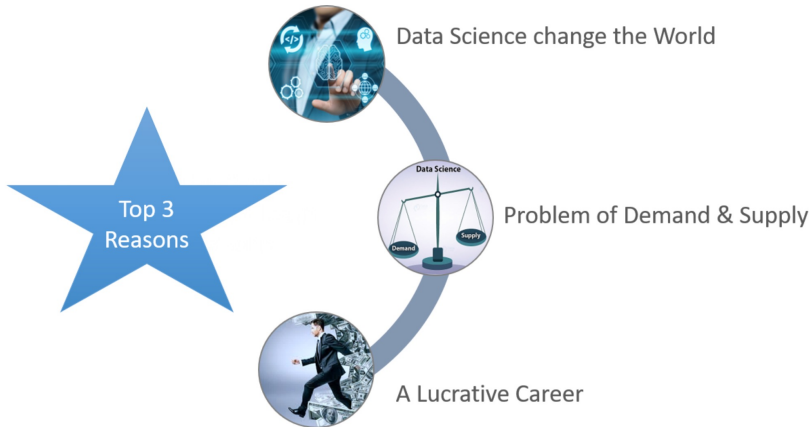
	Data scientist	Machine learning engineer	Data engineer
What they do	Build models that help business get better insights and make predictions from their data	Automate ML processes and make models work in a production environment.	Build, test and maintain data pipelines; provide ML models with quality data.
Skill set	<ul style="list-style-type: none"> <li>✓ Knowledge of math and statistics</li> <li>✓ Decision making and data optimization skills</li> <li>✓ High proficiency in SQL</li> <li>✓ Scripting skills (R/Python)</li> </ul>	<ul style="list-style-type: none"> <li>✓ Solid programming background</li> <li>✓ Data science skills</li> <li>✓ Knowledge of math and statistics</li> <li>✓ Rapid prototyping skills</li> <li>✓ Good problem-solving skills</li> <li>✓ Proficiency in deep learning frameworks</li> </ul>	<ul style="list-style-type: none"> <li>✓ Scripting skills (Linux commands)</li> <li>✓ Knowledge of databases</li> <li>✓ Knowledge of cloud technologies</li> <li>✓ Proficiency in SQL</li> <li>✓ Data modelling skills</li> <li>✓ ELT development skills</li> </ul>
Tools used	Python, R, Pandas, Jupyter notebooks, SQL	Python, PyTorch, TensorFlow, cloud services	SQL, Oracle, Hadoop, Amazon S3, Python

# Roadmap to DS





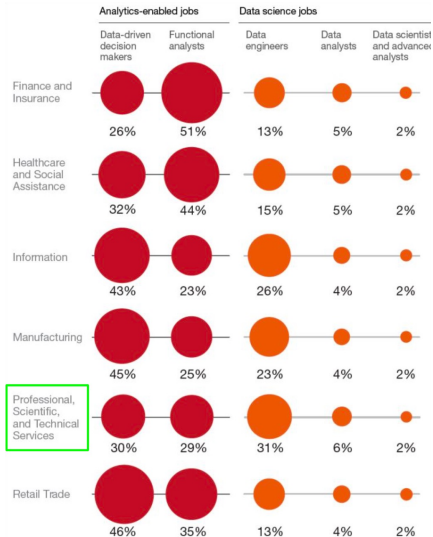
# Tại sao học khoa học dữ liệu?



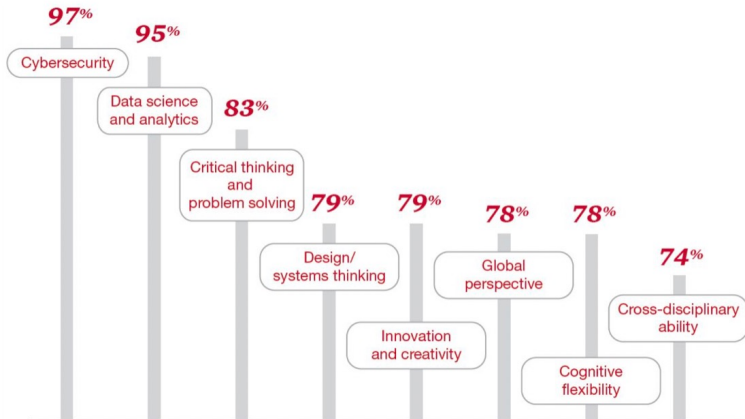
# Nhu cầu về các nghề nghiệp DS

- Xây dựng hệ thống dữ liệu
- Phân tích dữ liệu

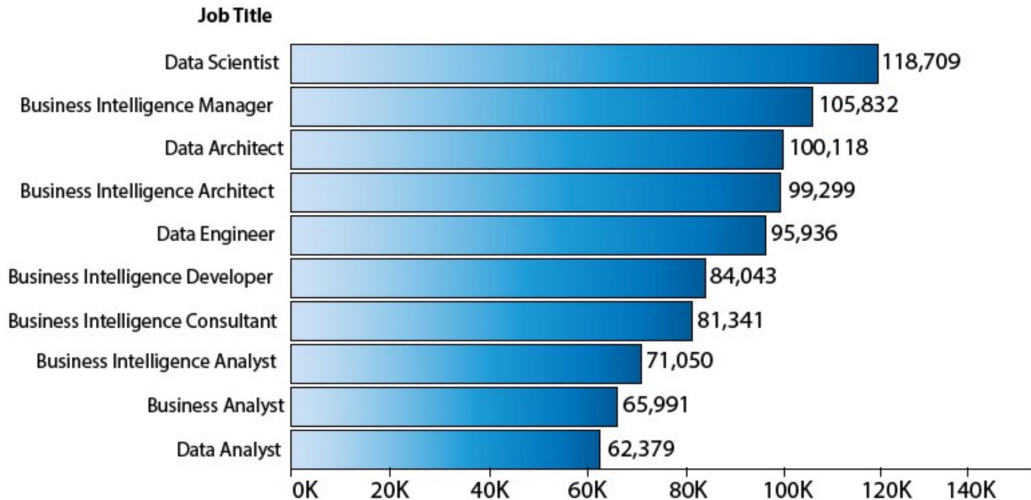
**Ở Việt Nam, nguồn cung nhân lực mới đủ 10% nhu cầu thị trường**



# Nhu cầu về các kỹ năng khó



# Mức lương hấp dẫn



Khoa học dữ liệu là gì

**Nội dung khoá học**

Thu thập và đọc dữ liệu

# Mục tiêu khoá học

Khoá học này cung cấp các kiến thức, kỹ năng cơ bản trong một dự án khoa học dữ liệu. Các chủ đề bao gồm:

- Các thao tác thu thập, đọc, làm sạch dữ liệu
- Khai phá dữ liệu sử dụng các công cụ thống kê
- Các bài toán hồi quy, phân lớp, phân cụm
- Các công cụ khoa học dữ liệu
- Ứng dụng khoa học dữ liệu ở một vài lĩnh vực nổi bật

# Kiến thức tiên quyết

Khoá học yêu cầu các kiến thức tiên quyết dưới đây. Sinh viên cần nắm vững các kiến thức này hoặc có khả năng tự bổ khuyết đầy đủ.

- Xác suất và Thống kê
- Cơ sở dữ liệu

Khoá học này sẽ lập trình trên ngôn ngữ Python. Sinh viên cần tự tìm hiểu Python trong quá trình học.

## Tài liệu tham khảo

Khoá học sử dụng 2 giáo trình miễn phí sau:

- *[ISL]: An Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, Laura Igual, Santi Seguí.
- *[ESL]: The Data Science Design Manual* tion. Seven S. Skiena

Các bài đọc mà giảng viên yêu cầu cần được hoàn thành trước mỗi buổi học.



# Lập trình và tính toán

- Sinh viên cần sở hữu máy tính để hoàn thành các bài tập hàng tuần (máy tính xách tay/để bàn tầm trung là đủ đáp ứng)
- Các bài tập sẽ lập trình trên môi trường Python 2.7.

Khoa học dữ liệu là gì

Nội dung khoá học

**Thu thập và đọc dữ liệu**

# Thu thập dữ liệu

- **Định nghĩa:** Là quá trình tập hợp và lưu trữ dữ liệu từ nhiều nguồn khác nhau để phục vụ xử lý các bước sau.
- Một vài nguồn dữ liệu phổ biến:



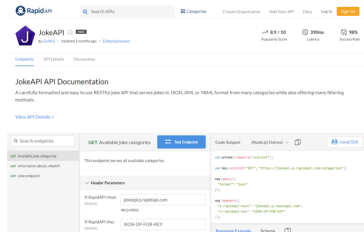
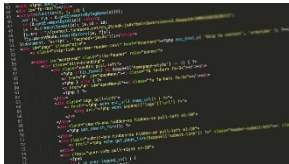
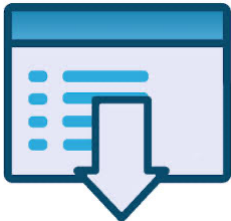
# Làm cách nào để thu thập dữ liệu?

## Ba câu hỏi cơ bản:

- Ai có dữ liệu?
- Họ có thể cung cấp cho ta với mục đích gì?
- Làm sao để truy cập được dữ liệu?

# Làm cách nào để thu thập dữ liệu?

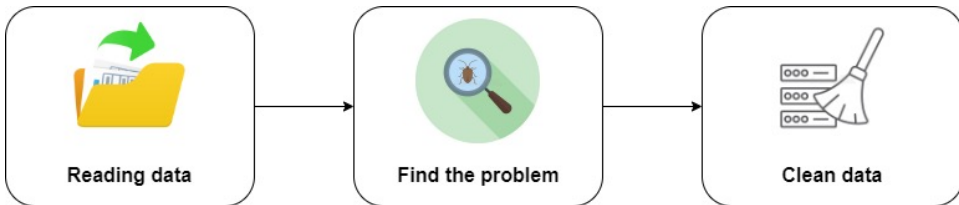
- Học phần này tập trung bàn về thu thập dữ liệu có sẵn trên mạng. Lý tưởng là tải trực tiếp được dữ liệu, còn không, hãy thử “cào” dữ liệu.
- Cào dữ liệu (Data Scraping / Crawling) là một phương pháp lấy dữ liệu từ các trang trực tuyến:
  - Cào thông qua HTML
  - Cào thông qua API



# Tiền xử lý dữ liệu

Là quy trình đảm bảo dữ liệu đủ tiêu chuẩn để phân tích và xây dựng mô hình.  
Tiền xử lý bao gồm:

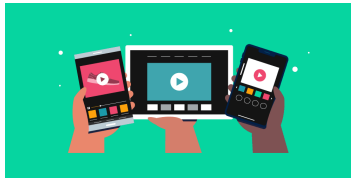
- Đọc dữ liệu
- Phát hiện vấn đề của dữ liệu
- Làm sạch dữ liệu



## Đọc dữ liệu

Khi đọc dữ liệu, cần quan tâm đến kiểu dữ liệu, vì nó sẽ quyết định cách thức xử lý. Một vài kiểu dữ liệu phổ biến:

- Dữ liệu dạng có cấu trúc như .csv, .xlsx(Excel), json,..
- Dữ liệu ảnh, video
- Dữ liệu văn bản



## Dữ liệu bảng

Dữ liệu dạng bảng gồm các cột và các dòng. Trong đó:

- Cột: Chỉ thuộc tính của bảng dữ liệu
- Dòng: Là một quan sát của bảng dữ liệu

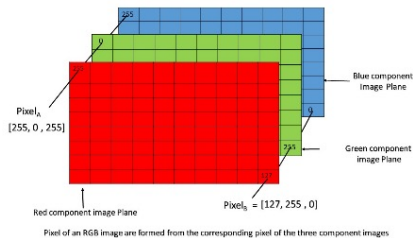
Phương pháp đọc dữ liệu:

- Khi dữ liệu là các file dạng bảng như .csv, .xlsx:
  - ☐ Sử dụng phần mềm Microsoft Excel
  - ☐ Sử dụng thư viện Pandas
- Khi dữ liệu nằm trong cơ sở dữ liệu:
  - ☐ Sử dụng phần mềm truy vấn cơ sở dữ liệu
  - ☐ Sử dụng thư viện Pandas



# Dữ liệu hình ảnh

- Mỗi hình ảnh coi như một bảng tạo thành từ các hàng và cột, trong đó mỗi ô của bảng ứng với một pixel.
- Một pixel lại coi như 1 vector 3 chiều ứng với 3 kênh màu Red, Green, Blue (*hệ thống màu RGB*)
- Phương pháp đọc dữ liệu hình ảnh:
  - Sử dụng phần mềm hiển thị ảnh
  - Sử dụng thư viện opencv (Python)



# Tổng kết

- Tổng quan khoa học dữ liệu: Định nghĩa, nhiệm vụ, quy trình
- Các nghề nghiệp khoa học dữ liệu và các kỹ năng khoa học dữ liệu
- Bước thu thập dữ liệu
- Bước tiền xử lý dữ liệu: Đọc dữ liệu